Role play with large language models

Year: 2023 | *Citations: 410* | *Authors: M. Shanahan, Kyle McDonell, Laria Reynolds*

Abstract

By casting large-language-model-based dialogue-agent behaviour in terms of role play, it is possible to describe dialogue-agent behaviour such as (apparent) deception and (apparent) self-awareness without misleadingly ascribing human characteristics to the models. As dialogue agents become increasingly human-like in their performance, we must develop effective ways to describe their behaviour in high-level terms without falling into the trap of anthropomorphism. Here we foreground the concept of role play. Casting dialogue-agent behaviour in terms of role play allows us to draw on familiar folk psychological terms, without ascribing human characteristics to language models that they in fact lack. Two important cases of dialogue-agent behaviour are addressed this way, namely, (apparent) deception and (apparent) self-awareness.